DOCUMENT RESUME

ED 252 560 TM 850 029

AUTHOR McArthur, David; Chou, Chih-Ping
TITLE Interpreting the Results of Diagnostic Testing: Some Statistics for Testing in Real Time. Methodology Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 84
GRANT NIE-G-84-0112-P1
NOTE 38p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Ability; *Adaptive Testing; Bayesian Statistics; *Diagnostic Tests; Hypothesis Testing; Latent Trait Theory; *Mathematical Models; Test Construction; *Test Interpretation; Test Theory

ABSTRACT

Diagnostic testing confronts several challenges at once, among which are issues of test interpretation and immediate modification of the test itself in response to the interpretation. Several methods are available for administering and evaluating a test in real-time, towards optimizing the examiner's chances of isolating a persistent pattern of erroneous performance by a student. Under ideal circumstances, a student who misunderstands the test content would be identified early in a testing sequence; from this point the test could be tailored to estimates not only of ability but also (or instead) to the relative likelihoods of a set of competing diagnostic hypotheses that could account for the student's behavior (ability). Items which could discriminate among these hypotheses could be administered in increasingly well-bounded subsets until a specified stopping rule is met. The following models for this procedure are described and compared: Wald's sequential probability ratio test, Sixtl's modified binomial method, Choppin's catenating Bayesian method, Fink and Galen's decision path method, Shortliffe and Buchanan's inexact reasoning method, Kmietowicz and Pearson's ranked probability method, and Schum's cascaded inference method. (BW)

DELIVERABLE - NOVEMBER 1984

METHODOLOGY PROJECT

Interpreting the Results of Diagnostic Testing: Some Statistics for Testing in Real Time

David McArthur
Project Director

Grant Number

NIE-G-84-0112, P1

Center for the Study of Evaluation

Graduation School of Education

The project presented or reported herein was supported pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

Interpreting the Results of Diagnostic Testing: Some Statistics for Testing in Real Time

by

David McArthur and Chih-Ping Chou

Introduction

Diagnostic testing in education, as in a variety of other fields, confronts several challenges at once, among which are issues of test interpretation and immediate modification of the test itself in response to the interpretation. This paper explores a set of methods for administering and evaluating a test in real-time, towards optimizing the examiner's chances of isolating a persistent pattern of erroneous performance by a student. What is expected from these methods? What does each method take into account in the testing process? How do they compare with each other?

For well over half a century the diagnostic value of interpreting a student's choice of a particular wrong answer to a test item has been appreciated (Pressey, 1926). Contemporary test specialists point to the measurement strength inherent in formulating tests for which the item distractors carry specific meanings for the appraisal of student abilities and disabilities (Roid & Haladyna, 1982). The rapid development of computer technology in the last decade has almost eliminated the practical restrictions on such testing. However, the overwhelming predilection continues in favor of correct/incorrect response scoring. The probative value of a wrong response -- that is, its significance for or against one or another of a set of plausible diagnostic hypotheses -- is totally obviated

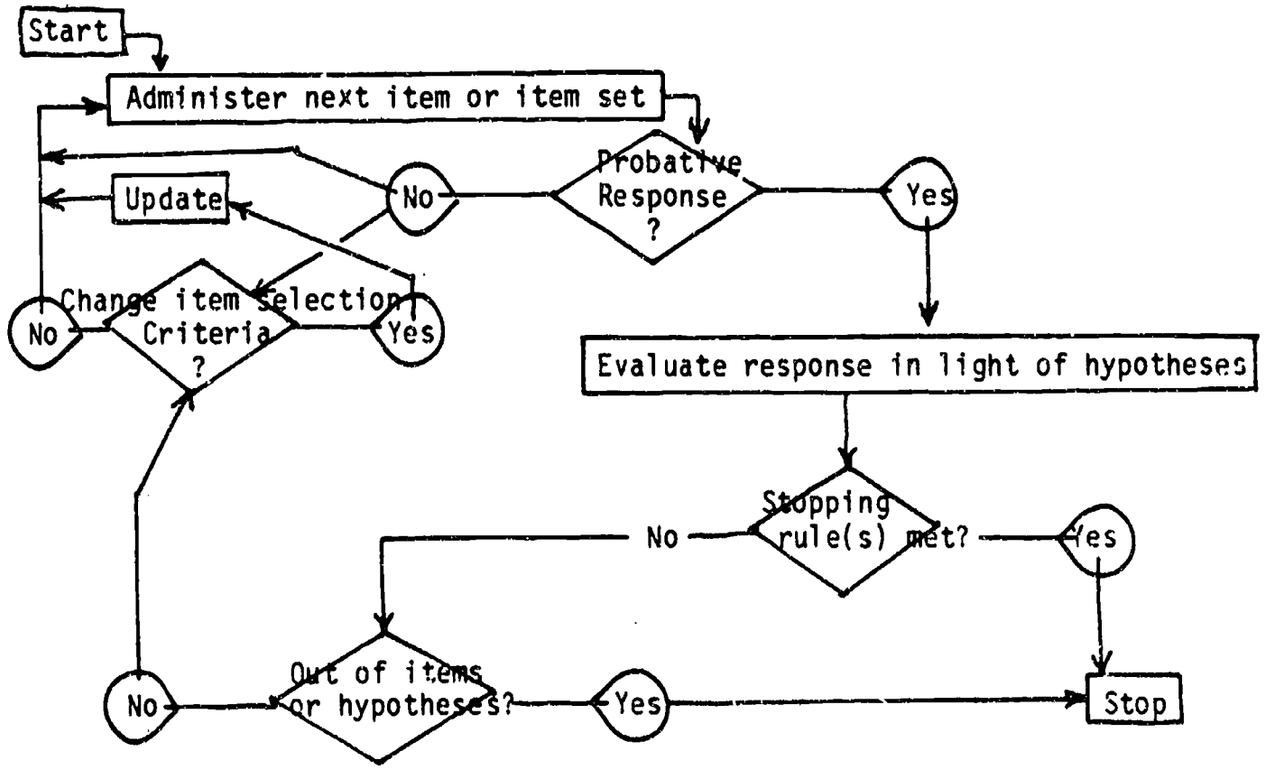
by conventional 0/1 scoring algorithms. Yet it is exactly that probative value which is central to forming diagnostic appraisals.

What is being sought in diagnostic testing is some cohesive pattern of wrong answers, a pattern of individual student responses which reveals a characteristic signature or diagnostic profile. Diagnostic profiles are an integral aspect of many psychological tests: a trained examiner probes with increasing selectivity and specificity until a meaningful psychological pattern appears. In recent work in projective testing, the initial response of the examinee to a stimulus card is codified, the ensuing inquiry is guided by estimates about the psychological dimensions of the problem as shown by that codifying, and the examinee's responses to that inquiry are used to refine and solidify one or another diagnostic inference (McArthur & Roberts, 1982). This honing procedure proceeds in an adaptive sequence based in part on technical guidelines, in part on the examinee's consistency (or lack thereof) in responding to the stimulus, and in part on the examiner's inferences of the strength of the present evidence and the benefit of continued testing.

Under highly idealized circumstances, the disability of a student who is engaging consistently in a certain misunderstanding of the test content would be identified early in a testing sequence by the astute observer (human or computer); from this point the test then could be tailored to estimates not only of ability (Θ) but also (or perhaps instead) the relative likelihoods of a set of competing diagnostic hypotheses $\{H_1, H_2, H_3 \ldots\}$. Items whose distractors would assist in discriminating among the plausible competing H's for that student's behavior could be administered to the student in increasingly well-bounded subsets until one or another

stopping rule is met. Briefly, the optimal stopping rule would be one which maximizes the likelihood of a single primary diagnostic hypothesis, supported by sufficient estimation strategies and by exactly the right amount of evidence. The evidence is not so much as to be unnecessarily redundant, and not so little as to be insufficiently discriminatory, not so difficult that the student simply flounders and not so easy that the examiner misses the problem altogether. This task is by its nature a compound probabilistic undertaking, although the flow chart which schematically illustrates this task is relatively simple (see Figure 1).

Figure 1
Schematic flow of a generalized response - contingent test



The flow of a response-contingent test is governed by two implicit prerequisites. The first is that a finite set of suitable hypotheses is represented by the test. The hypotheses are appropriate to age-level, intellectual functioning and motor capabilities of the target student. The hypotheses are orderly, in the sense that they are either at a uniform level of abstraction and mutually independent, or they fit an explicit hierarchy or cascade and are mutually dependent upon one another. The second implicit prerequisite is that a given item or set of items be closely linked to at least two competing hypotheses. A response must be able to be evaluated in terms which tie the response to one hypothesis but mismatch another; the response cannot be considered probative unless these links can be made at the time the response is given.

As the test is administered four decisions must be made in real-time. 1, Is the response probative? If not probative, further decisions regarding discrimination among competing hypotheses are obviously moot; questions must be asked as to the appropriateness of the item given, item selection criteria, and for the original hypotheses, then another item or item set readministered. 2) Is any one of the stopping rules in use met? If a stopping rule applies, it signifies that the examiner has reached an applicable criterion, so further testing is not warranted.[1] 3) Are there remaining items to administer, or remaining hypotheses for which one or more stopping rules have not been met? If either answer is negative, there is nothing to be gained by further testing in the context of the present

[1] This assumption holds if the examiner considers stopping rules disjunctive. If stopping rules are considered conjunctive, then the question is answered in the negative until all associated stopping rules are met -- with, of course, a larger volume of responses and presumably, though not necessarily, an increased discriminative power.

test. The presence or absence of one or more supported hypotheses, and the costs of continued testing with additional instruments, govern the examiner's decision at this juncture. 4) Should item selection criteria be changed? If no hypothesis is supported, and if a bank of items and hypotheses remain, a decision must be made as to whether the sequence of administration continues to be appropriate. Explicit branching can occur here; interactive tests use this decision point to change topics, item complexity, and/or task requirements to enhance the expected likelihood of hypothesis discrimination. It is this decision point which allows the examiner to maximize inferences in regard to diagnostic hypotheses.

With very rare and specialized exceptions, diagnostic testing in education seldom enables the test interpreter to build on inferential strategies with respect to individual test performance. Moreover, a variety of theoretical and practical problems appear to have plagued developments along this line. Among the problems that arise in the pursuit of interpretable patterns is the difficulty in obtaining diagnostic performance clusters from raw data without a prior set of likelihood estimates for a small and workable number of competing hypotheses.[1]

The problem of assigning meaning to cohesive patterns of response reduces in its simplest form to two elements: limiting the number of observations we need to take, and limiting the number of possible

---

[1] The number of possible clusters m which can be made out of n observations is a Stirling number:

$$S_n^m = \frac{1}{m!} \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} j^n$$

Unfortunately, even for a handful of observations, this term can be exceedingly large.

meaningful clusters into which we will place observations. A test ought to provide enough range to evaluate fairly a highly varied set of possible examinees, without building such a long test that any of the protagonists -- examinee, test administrator, test interpreter, or test designer -- is exhausted by the process. The diagnostic process ought to involve checking in real-time as to whether any payoff remains for administering more test items. Is the performance of the examinee at this moment in time sufficient in quantity and "cohesion" for us to draw a suitable diagnostic inference?

A simple stopping rule for diagnosis takes the following form: go no further because any one of several probabilistic boundaries is met. Among the set of allowable hypotheses, one diagnostic hypothesis has emerged in the "lead." One possibility for limiting observations and limiting clusters simultaneously is to avoid that Stirling number by picking an easy criterion, a low threshold of confidence, and a small number of allowable hypotheses. Alternatively, we can limit the number of possible clusters for diagnosis to exactly two, so students must select one option or the other; the stopping rule becomes: go no further when one hypothesis obtains a simple majority of examinee responses.

Foremost among the difficulties of using the stopping rule approach to limit observations and clusters is the extreme paucity of situations in educational or psychological testing for which a strict parsimony of hypotheses can be formulated. Another is the reasonable assurance that some students will guess some of the time on some items. Yet another is the degree of confidence one places in a single response as a marker of a general pattern of responses; a test item, after all, is seldom adequate as

a mirror of a student's true understanding. Other difficulties arise in regard to assessment of the several probabilities that contribute to the flow of the test: they include problems with probabilistic comparison baserates, fuzziness in the Bayesian priors, and inherent objections to traditional Bayesian probabilistic analysis itself.

None of the problems stated here is insurmountable. Theoretically useful probabilistic algorithms for diagnostic inference are found in several professions. This paper sets out six algorithms which have bearing on the interpretation of response patterns and diagnosis. Two are drawn from probabilistic methods in educational testing -- Sixtl's modified binomial and Choppin's catenating Bayesian methods; two are drawn from recent developments in medical diagnostic studies -- Fink and Galen's decision path analysis and Shortliffe's inexact reasoning; one rests in decision theoretic analysis -- Kmietowicz's ranked probabilities; and one builds on a Baconian probabilistic appraisal -- Schum's "cascaded inference," which has been studied primarily in the context of decision making in jurisprudence. Each of the six methods will be placed into a common notation, and a comparison made between the advantages and disadvantages of each, with special attention to the restrictive nature of prerequisites and the relative strength of the stopping rules. Not all of these approaches are equivalent in scope, nor do they have analogous assumptions about the patterns which are being isolated from the raw data. It is also important to note at the onset that the stronger inferential procedures inevitably impose more restrictive conditions on the user.

ESSENTIALS FOR PROBABILISTIC EVALUATION OF DIAGNOSES

In the present discussion, the following terms are used throughout:

$\{H\}$	the set of alternative hypotheses $\{H_1, H_2, H_3 \ldots H_m\}$ (includes $H_{correct}$)
H_i	the i^{th} hypothesis of $\{H\}$, contained in one or more alternatives for one or more items
$P(H_i)$	the prior probability of H_i
x_s	the examinee's response at a given step s in the testing sequence (generally a response to a single item which represents a choice of H_i, from the set of $\{H\}$).
H_j	those hypotheses shown to the examinee in an item but not selected
$H_{\bar{j}}$	those hypotheses not shown to the examinee in the item, so not selectable at this step
k	the number of hypotheses contained in an item's answer choices
m	the number of hypotheses in all ($m \geq k$)
n	the number of attempts made by the examinee ($n \geq x$)

Diagnostic testing involves several key terms made up of the above entries. The general form of the stopping rule is the following: At a given step s in the sequence of the test, does the accumulated evidence $\{X+\}$ which suggests H_i exceed the accumulated evidence $\{X-\}$ which relates instead to H_j and $H_{\bar{j}}$? The accumulation of evidence on both sides is treated probabilistically, and the likelihood ratio that results from dividing one into the other is assessed against an allowable lower and upper limit. Wald's (1947) sequential probability ratio test (SPRT) is the earliest treatment of this stopping question:

$$\Lambda = \prod_{i=1}^{n} \frac{P(\{X\}|H_i)}{P(\{X\}|H_j)}; \quad \text{lower limit} < \Lambda < \text{upper limit} \qquad (1)$$

A number of studies have applied Wald's (1947) sequential probability ratio test to the task of test individualization (cf. Ferguson, 1969, 1973). The SPRT, predicated on Bayesian methodology, is well-understood but clearly does not begin to account for the variety of factors which contribute to examinee performance.

Gorry and Barnett (1968) showed that sequential diagnostic testing involves a compounding of conditional probabilities as follows:

$$P(H_i|x) = \frac{P(x|H_i \cap \{X\})\, P(H_i|\{X\})}{\Sigma P(x|H_j \cap \{X\})\, P(H_j|\{X\})}$$

where $\cap$ implies a logical "and" using the entire set of behaviors acquired to date $\{X\}$ as evidence. To be successful, these approaches require extensive knowledge of prior conditionals and interrelationships. They are fairly impractical except in highly controlled environments. The various techniques which follow are to be viewed as approximations of these data-intensive methods.

TECHNIQUES FROM EDUCATIONAL RESEARCH

Modified Binomial Method

A decade ago a German educational researcher published a paper on automated test administration which included an application of Bayesian probabilistic analysis with correction for guessing. Sixtl (1974) presented a formula for a stopping rule which acknowledges the roles of item answer alternatives, and is readily identified as a classical Bayesian approach to forming a likehood function for a particular hypothesis.

Sixtl's likelihood ratio, when modified to cover each diagnostic hypotheses of a set of hypotheses, reads as follows:

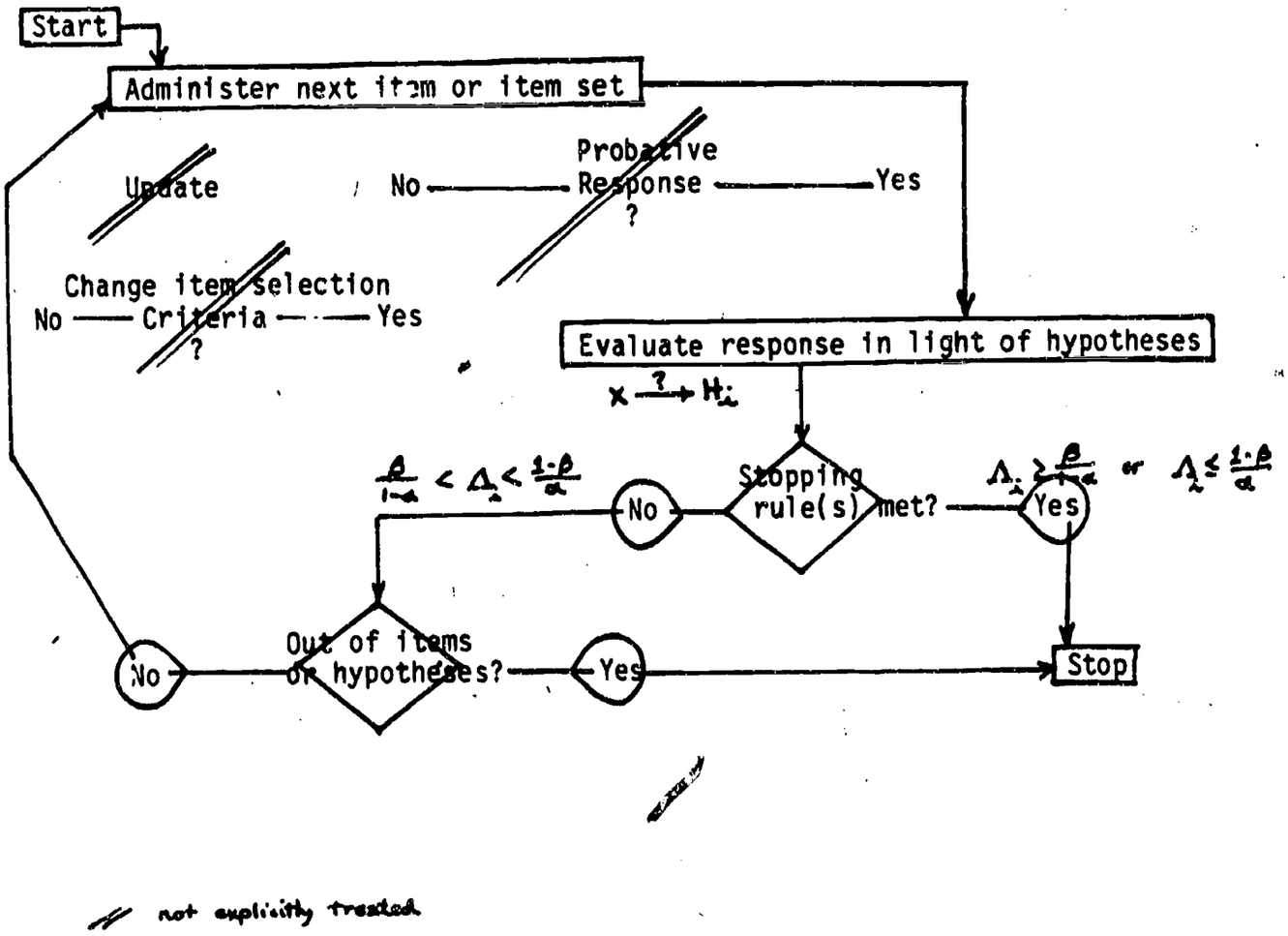
$$\underset{i \in \{H\}}{\Lambda} = \frac{\left[\frac{k-1}{k}\, p(H_i) + \frac{1}{k}\right]^{x_i}}{\frac{1}{k}} \cdot \frac{\left[1 - \left(\frac{k-1}{k}\, p(H_i) + \frac{1}{k}\right)\right]^{n-x_i}}{\frac{1}{k}} \qquad (2)$$

Sixtl's approach involves selection of a Bayesian prior for each diagnostic hypothesis, involing a simple correction for guessing, and construction in real-time of the likelihood function Λ for the hypothesis to fit the stopping rule

$$\frac{\beta}{1-\alpha} < \Lambda < \frac{1-\beta}{\alpha}$$

where α and β are conventional measures of significance and power, respectively. Figure 2 presents a schematic illustration of this method.

Figure 2
Sequential testing - binomial model,
multiple hypotheses in {H}



Immediate objections can be made to Sixtl's approach. First, the model of guessing is simplistic; it allows only a constant term for a function that is unlikely to be stable across items and respondents. Second, the fixed nature of the Bayesian priors must be chosen to reflect {H} without regard to context or sequence effects. Third, Sixtl's procedure fails to use all of the information gained at a given moment to form an updated chain of hypothesis evaluation.

Catenating Bayesian Method

A sequential system for response contingent diagnostic testing was proposed by Choppin (McArthur & Choppin,1983) using both Bayesian priors and conditionals to form a continuously updated probability assessments for diagnostic hypotheses in real time. Choppin's approach, modified slightly to reflect iterative cycles through $\{H\}$, is:

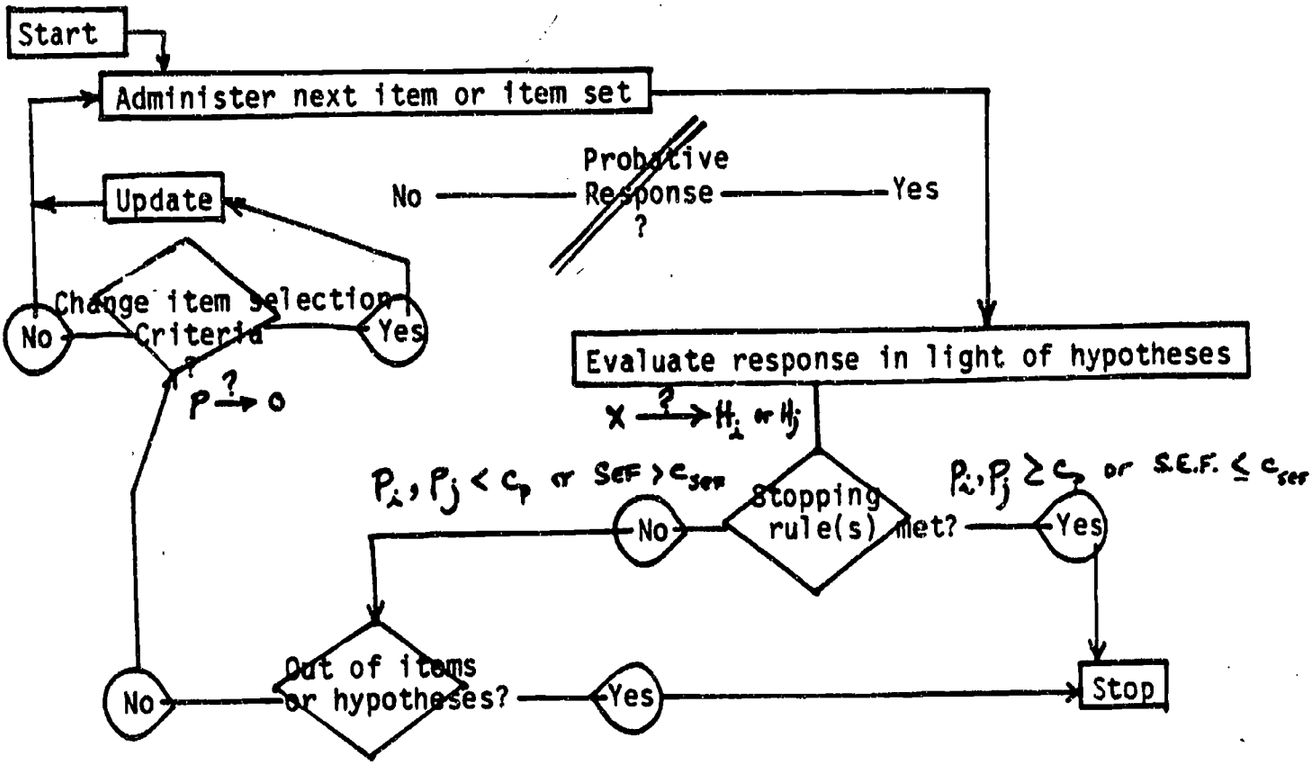
$$P_{i_n} = \frac{P_{i_{(n-1)}} \; p(X_i \mid H_i)}{\sum_{\substack{i \in \{H\} \\ j \neq i}} P_{j_{(n-1)}} \; p(X_i \mid H_j)} \qquad (3)$$

The computation is predicated on a catenating sequencing of conditionals: initially it requires priors for X_i and X_j assuming H_i is true. Each is updated upon the examinee's next selection, such that Choppin's P in Bayesian terms is a catenating conditional ratio appraisal. Use of the Shannon entropy function (Gleser,1974), which in this context is

$$S.E.F. = -\sum_i P_i \log P_i \qquad (4)$$

simplifies the output of the catenating method by concluding at each step with a single expression for the uncertainty remaining in the set of proportions. The largest decline in S.E.F. denotes an optimal stopping for the sequence; the largest p(H) at that step is taken as an optimal H for that respondent. Figure 3 shows this procedure at work.

Figure 3
Sequential of testing -- catenated Bayesian model, multiple hypotheses in $\{H\}$



// not explicitly treated ; C: constant

An inclusive conditional probability $P(x_i|H_j)$ represents the probability that selection i would be made when H_j is true, a selection which could be made for a wide variety of reasons. One immediate objection to Choppin's P is that the catenation is sensitive to the choice of the separate initial priors. A subtle but potentially damaging argument is also to be found in the catenation and recomputation of conditionals under $H_{\not j}$, when an alternative hypothesis $H_{\not j}$ is not represented among the

item distractors. Unfortunately, on both counts it is the Bayesian system of probabilistic assessment itself which forces this to occur.

TECHNIQUES FOR DIAGNOSTIC RESEARCH FROM OTHER AREAS

Decision Path Method

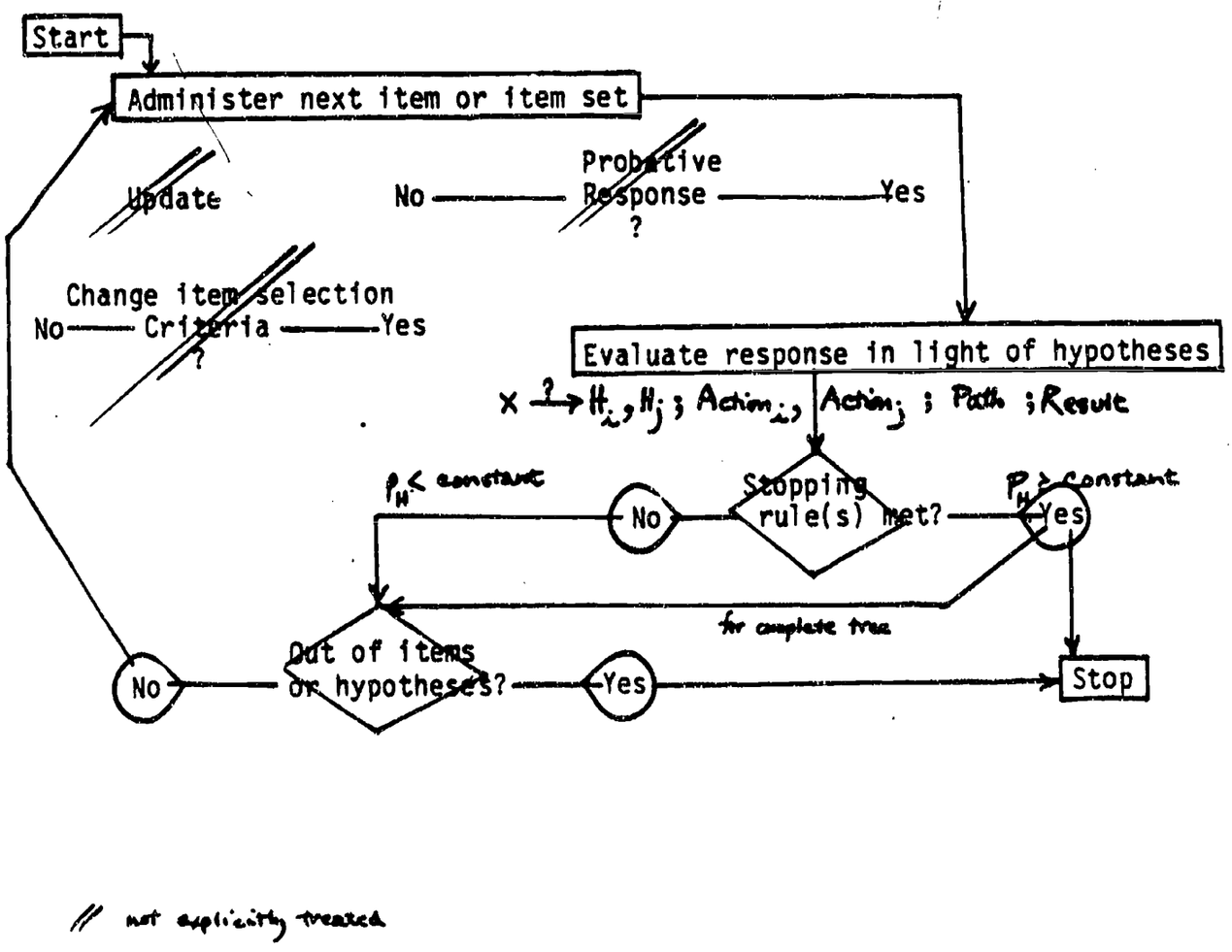
At any point beyond the entry point in a sequential test, an additional set of conditional probabilities which are potentially important are required. Not only are there conditional interrelationships among the $\{H\}$ and $\{X\}$, but also among the paths which led up to the particular step in the test sequence and the actions taken by the examiner at each step. An applied extension of Bayesian analysis to decision paths is found in the field of research in diagnostic medicine The decision tree analysis illustrated by Fink and Galen (1981) invokes a Bayesian framework operating with compound conditionals:

$$p(x \mid \{H\}, \{A\}, Path_k, R) = \frac{p(R_i \mid x, \{A\}, Path_k) \cdot p(x \mid \{A\}, Path_k)}{\sum p(R_j \mid x, \{A\}, Path_k) \cdot p(x \mid \{A\}, Path_k)}$$

where $\{H\}$ = the hypotheses allowable within a given situation, $\{A\}$ = the actions to be taken within the situation, $Path_k$ = the path from preceding selections which led to the current condition, and R the result of selecting a particular action. This result leads to further data which then allows refinement of the probability estimate for H_i. The multiple conditioning terms lead to extensively annotated decision trees, for which information is available about the relative values of selecting one option

over another in terms which includes the sequence of those options, their cost, their efficiency, and their measurement certainty. Figure 4 illustrates the calculation for the decision tree method schematically.

Figure 4
Sequence of testing - decision tree model,
multiple hypotheses in {H}



// not explicitly treated

An obvious implication of Bayesian path analysis is that, when given fully elaborated baserates, a researcher can construct a fully elaborated decision tree which includes each possible diagnosis, all possible interactions among diagnoses, and cost-efficiency assessments. An obvious

hitch in applying the system to educational testing is the profound lack of reliable baserate data for all but the least complex diagnostic hypotheses likely to be explored. Additionally, distinctions between various paths may be far less profound in the context of educational testing than in diagnostic medicine.

Inexact Reasoning Method

In a variety of settings, evidence about prior probabilities is relatively limited. If the priors can be estimated, we can draw on a system for hypothesis evaluation called the method of inexact reasoning, which accounts for the lack of exactitude in the establishment of priors. It was developed by Shortliffe and Buchanan (1975) in the context of the well-known MYCIN automated medical diagnosis program. Its prime concern is with the strength of evidence, rather than a perfect match between evidence or behavior and hypotheses. Three separate terms are required: one a measure of belief and another a measure of disbelief, expressed as conditional statements, plus a term which reflects the difference between belief and disbelief:

$$MB_{H_i} = \begin{cases} 1 & \text{if } p(H_i)=1; \\ \dfrac{\max[p(H_i|x),\ p(H_i)] - p(H_i)}{\max[1,0] - p(H_i)} & \text{otherwise.} \end{cases}$$

$$MD_{H_i} = \begin{cases} 1 & \text{if } p(H_i)=0; \\ \dfrac{\min[p(H_i|x),\ p(H_i)] - p(H_i)}{\min[1,0] - p(H_i)} & \text{otherwise.} \end{cases}$$

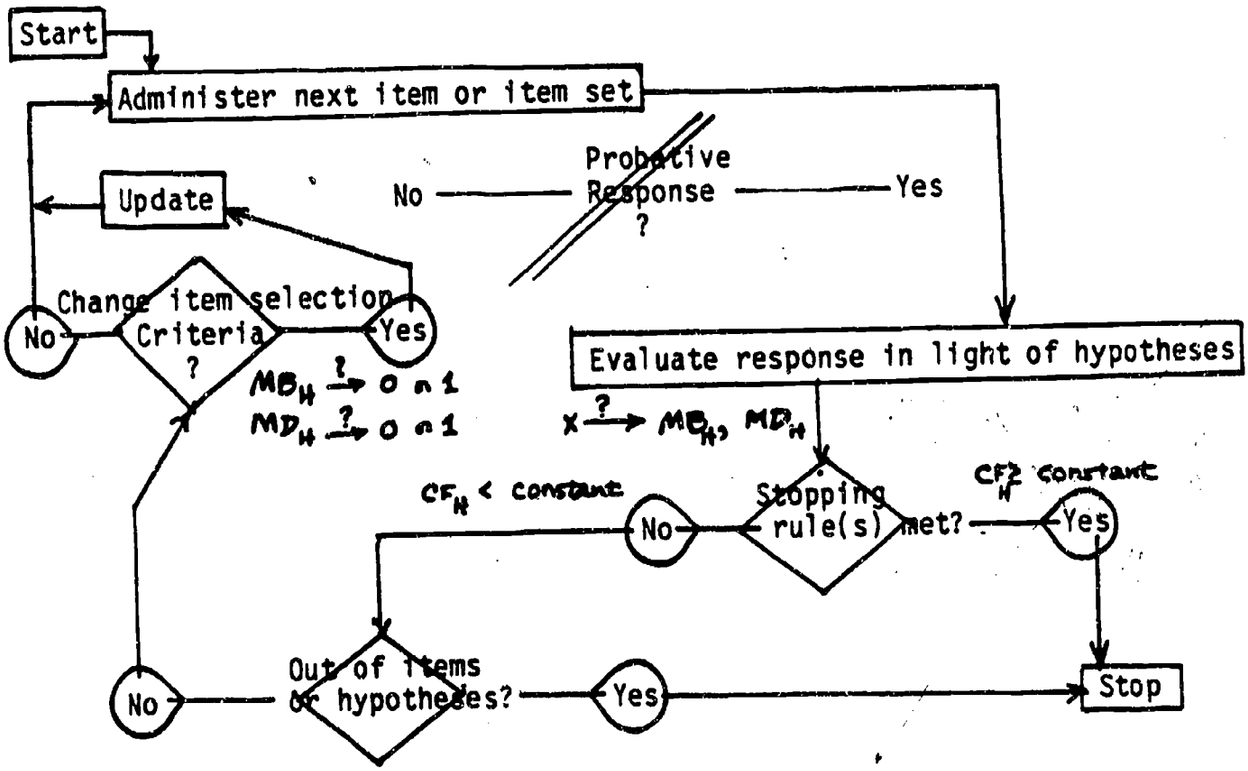
$$CF_{H_i} = MB_{H_i} - MD_{H_i} \qquad (5)$$

In many ways the notation appears to more closely reflect the psychological mindsets and inductive decision processes used by practicing clinicans than

the preceding methods, which are formally more exact (an important point discussed later in this paper).

Originally this model was put forward as a system of approximating conditional probabilities, suitable for circumstances characterized by data which tends to be subjective rather than objective. Since very few of the natural sciences have exact data in the strict sense required by Bayesian conditionals, the reasonableness of pursuing approximations seems assured. Moreover, many outcomes of a decision process are not even at the same level of rough granularity as the data used in that process; that is, the number of remedial options available to an examiner are fewer than the number of diagnostic clusters for performance of an examinee. Thus an approximation, if adequate, can provide completely sufficient guidance to the examiner for the purposes at hand. At minimum the approximation should provide a basis for corroborating human judgments of logical premises, actions, and consequences. Indeed, the model was incorporated into a highly regarded artificial intelligence approach to medical decision making which itself has seen extensive development and generalization. Figure 5 illustrates this approach at work.

Figure 5
Sequence of testing - inexact reasoning model,
multiple hypotheses in $\{H\}$



// not explicitly treated

Problems with the approach are unavoidable. Adams (1976) elaborated a series of theoretical objections which focus on the direct relations -- not immediately obvious -- between MB, MD, CF, and conventional Bayesian solutions to serially adjusted probabilities. Again, because of Bayesian logic, one can rapidly arrive by computation at untenably small conditional probabilities even when intuitive logic suggests otherwise. The strongest theoretical failing lies in the assumption of independence of $\{H\}$; any interdependence goes unaccounted in MB, MD, CF. As CF constitutes a

weighting factor its role in practical applications of MB and MD is multiplicative, but, Adams claims, "not true in general."

> The fact that in trying to create an alternative to probability theory or reasoning Shortliffe and Buchanan duplicated the use of standard theory demonstrates the difficulty of creating a useful and internally consistent system which is not isomorphic to a portion of probability theory (p. 185).

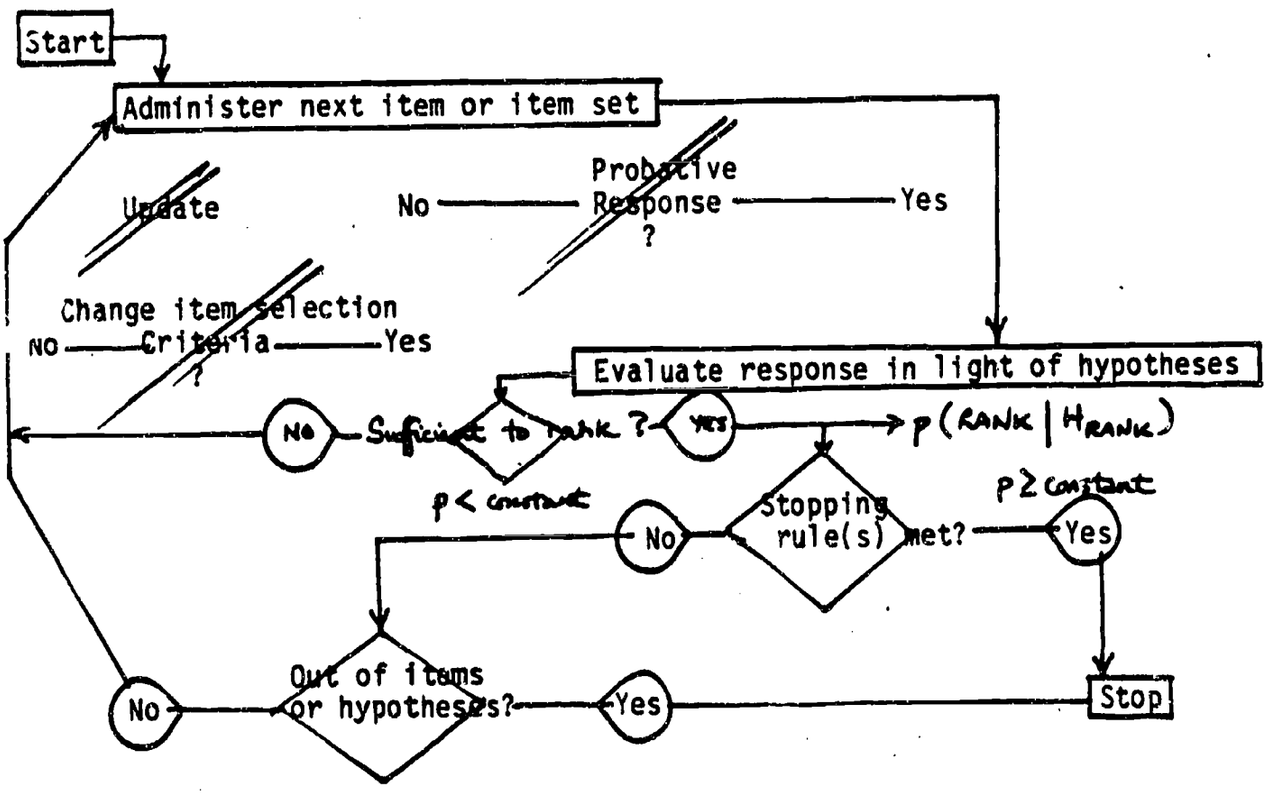
Ranked Probability Method

At the lowest end of the spectrum in terms of conditional complexity is a method which requires no more than weakly ordered priors of the form $p(H_i) > p(H_{i+1})$. In a variety of settings the researcher labors with unknown (and potentially unknowable) data about which only a minimum degree of information can be stated with confidence. In such conditions of incomplete knowledge, Kmietowicz and Pearson (1981) have spelled out a decision theory, and Horbar (1983) has illustrated an application to medical diagnosis. Using Horbar's approach, we can state the following:

$$\text{RANKING:} \quad p(x|H_1) \geq p(x|H_2) \geq p(x|H_3) \ldots \quad \begin{cases} \geq \text{weak ranking} \\ > \text{strong ranking} \end{cases} \qquad (6)$$

From a series of tables, generated by a procedure involving random sets of priors and conditionals, the user determines the probability that a given ordering of $\{H\}$ shown by the examinee's responses reflects an expected ordering. For example, the order $H_3>H_2>H_1$ has a substantially smaller posterior probability in reference to the expected sequence $H_1>H_2>H_3$ than does the order $H_2>H_1>H_3$. Figure 6 illustrates this approach at work in a

Figure 6
Sequence of testing - ranked probability model,
ranking of multiple hypotheses



// not explicitly treated

hypothetical testing situation.

At its heart the ranked probability method is a Bayesian procedure with a loosening of terms. It assumes that the examiner can reasonably generate an expected sequence for $\{H\}$; it also assumes that the elements of $\{H\}$ are mutually exclusive. Of concern here is that the tables themselves may be contingent in important ways on the original procedure which produced them (Horbar, personal communications). Additionally, no account is made of reliability of the evidence or of guessing behaviors and other

nonrandom choices by the examinee. A great deal of work has been produced in the area of decision theory, but extensions of such methods to situations involving incomplete knowledge are very scarce. An alternative system which addresses incompleteness mathematically is found in a terse monograph by Vesely and Vajda (1971). Further developments are essential.

Cascaded Inference Method

In situations with conflicting evidence such as are likely to be generated by a diagnostic test, it would be exceedingly helpful to have a system of analysis which takes account of the conflict and in particular the degree to which a given item response x relates to discrimination among the set of diagnostic hypotheses $\{H\}$.

For obvious reasons, the problem of developing conclusions from bits of evidence -- some corroborating, others contradictory, some useful, other useless, some fresh, others redundant -- has been of interest to researchers in jurisprudence. In the typical setting, a jury faces multiple and deliberately conflicting sources of information -- testimony by witnesses for the prosecution and the defense, documentation, photography, statements by court and counsel, and must develop a collective judgment as to a binary $\{H\}$ consisting of "guilty, not guilty." The Bayesian system of mathematical logic collapses under the demands of inferential reasoning required here; for example:

> ...testimony requires [a jury member] to assess the likelihood that the defendent was, in fact, at the scene/time. This foundational stage involves evaluation of the witness's credibility. Then, assuming the defendant at the scene/time of the crime, one must assess how strongly this event bears on the issue of whether or not the defendent committed the crime. Further difficulty is presented by intricate patterns of reasoning which require the joint consideration of current evidence with one or more previously given piece of evidence (Schum & Martin, 1982, p.106).

Schum draws on a Baconian approach to inductive reasoning explicated by L. Jonathan Cohen (1977, 1982) which allows direct estimations of probabilities for inference structures. Inference structures, which are found in all forms of human reasoning, run as follows:

> I have an assertion about x, which I read with some degree of skepticism, and which I take as a reflection on facts or events which in time I combine to assess the "major or ultimate facts-at-issue."

In a jury setting, a witness gives testimony about the crime that occurred. It may consist of an event which can be linked directly to guilt or innocence of the defendent ("I saw him rob the lady"). Such first-order relations of x to $\{H\}$ are remarkable because they are so rare. Witness testimony is more often of a fact that may or may not be interpreted as probative of events which may or may not be linked incompletely to guilt or innocence ("I heard a scream and saw someone running"). These compound cascades of inference to facts-at-issue are represented by extensions of the likelihood ratio

$$\underset{i \in \{H\}}{\Lambda}_{j \neq i} = \frac{P(x \mid H_i)\left[P(x \mid \{X+\}, H_i) - P(x \mid \{X-\}, H_i)\right] + P(x \mid \{X-\}, H_i)}{\sum P(x \mid H_j)\left[P(x \mid \{X+\}, H_j) - P(x \mid \{X-\}, H_j)\right] + P(x \mid \{X-\}, H_j)}$$

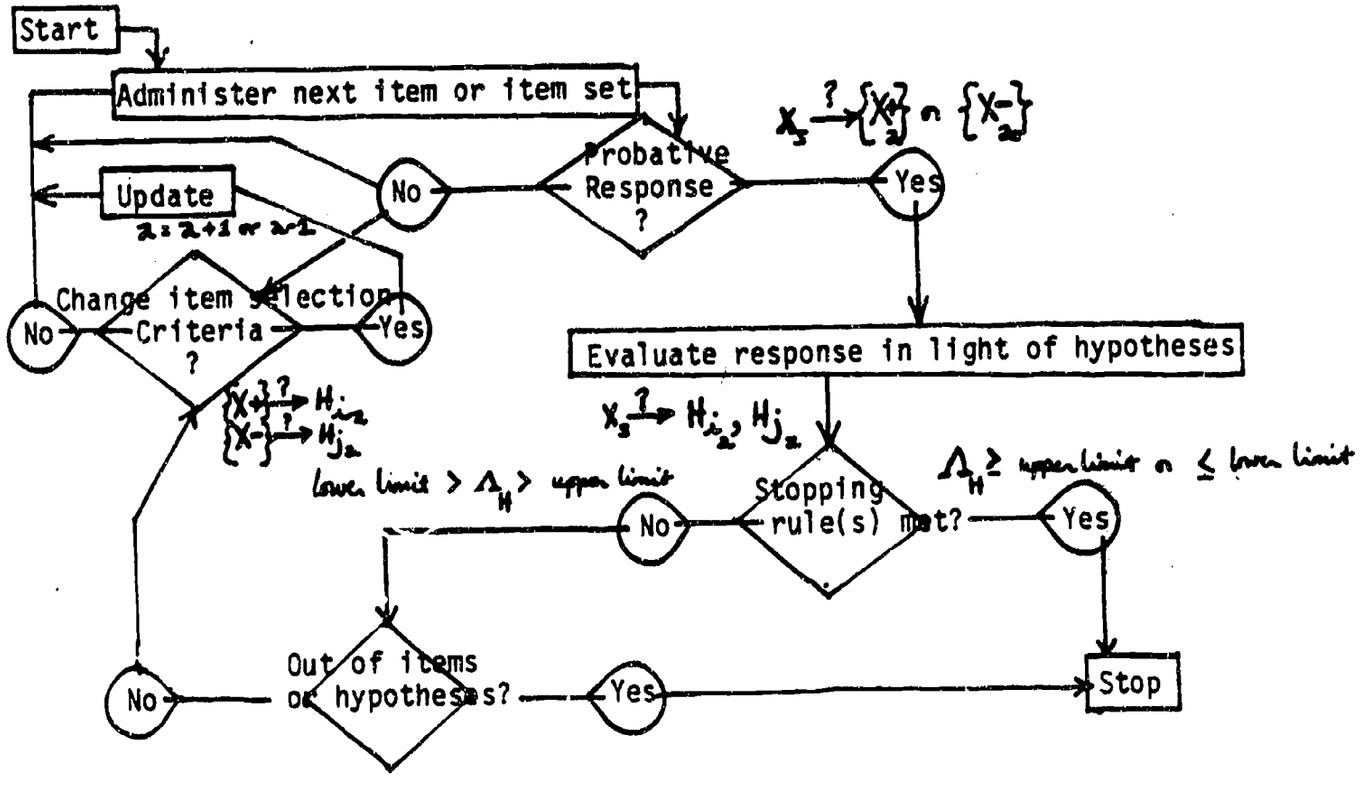
where $\{X\}$ is the set of evidence accrued to date, $\{X+\}$ is the portion of that evidence in corroboration with the testimony to date, $\{X-\}$ is the portion of that evidence in contradiction to the testimony to date.

The cascaded likelihood ratio has interesting properties, notably the use of terms which speak to the contribution of x to the set of evidence $\{X\}$ pointing to H_i. The first of these resides in the term which contrasts the relation of x to $(X \& H_i)$ minus the relation of x to

$(X-, H_j)$, in comparison to the relation of x to $(X+, H_j)$ minus the relation of x to $(X-, H_j)$. These elements relate the degree of specificity of x to H_i. The selectivity of x to $\{X-, (H_i \text{ or } H_j)\}$ is an expression of how distinctly x discriminates itself from the portion of $\{X\}$ which is contradicted by x. In more familiar language, the combinations formed here address true positives, false positives, true negatives, and false negatives.

Figure 7 is a demonstration of cascaded inference at work. In the

Figure 7
Sequence of testing - cascaded inference model,
multiple hypotheses in $\{H\}$



s = step ; a = inferential level

figure, the individual elements of evidence are annotated by a subscript indicating when the evidence is acquired. The full complexity of the calculations is not shown.

The analogy of cascaded influence in jurisprudence to cascaded inference in diagnostic testing can be seen in the following example. A student behaves in a fashion scored by x, which while not to be taken as a direct assertion of H_i, is seen as a symptom of H_i and thus an element of confirmation to $\{X+, H_i\}$, and an element in disconfirmation to $\{X-, H_j\}$. A student's erroneous response to a math test item is scored as symptomatic of a logical misunderstanding of how to carry digits in two-digit subtraction: the examiner includes this response in forming an overview of that student's pattern of responses across the test, but weights this response by

- the degree to which the response is probative for H_i
- the degree to which the response in consistent in X+ and $\{X, H_i\}$
- the degree to which the response is contradictory to X+ and $\{X, H_i\}$

Because of Baconian probability techniques, a relatively low-frequency response may contribute effectively to discriminating among $\{H\}$, and to directing the examiner to choosing a suitable item which may also have low (though nonzero) hypothesis likelihoods.

The underlying logic of the cascaded inference model and its explication of inference structures appears to closely resemble the logic and inference structures used by juries. By extrapolation, the same logic and inference structures describe the task of an educational or psychological diagnostician. At the present time, however the cascaded inferences model has not been tried with educational test data.

Comparison of techniques

The six analytic schemes presented thus far are chosen to reflect a series of constrast in assumptions, prerequisites, processes, and outcomes. The sequence portrayed above is an attempt to let each new method address the failings of the method that preceded. To begin, Sixtl's modification of the sequential probability ratio test shows an accounting for the probability of responding by chance. Choppin's catenating technique incorporates conditional probabilities beyond the single $p(H_i)$ used by Sixtl; these allow one to chain together the evidence of $p(H_i, H_j, H_k)$. Fink and Galen's decision tree moves from unconditional priors to compound conditional priors of the form $p(X|A,B,...)$ where $A,B...$ represent elements of the context surrounding the observation x -- that is, what path was used to arrive at this observation, what action was taken, and so forth. The ranking method of Kmietowicz, put into practical terms by Horbar, is in theory a relaxation of requirements; where the decision tree method requires a great deal of hard evidence, the ranking method can make use of knowledge about unconditional prior probabilities that is much less complete. The inexact reasoning method of Shortliffe and Buchanan attempts to portray both unconditional prior and conditional estimates of probability in a system that also loosens the need for exact or strictly ordered data. The cascaded inference method, developed from the work of Cohen by Schum and colleagues, attempts to correct the restrictions of Bayesian probabilistic reasoning, to allow hierarchical and nested hypothesis evaluation.

As noted at the outset, the various techniques differ markedly in their requisite assumptions and scope. Figure 8 presents a listing of

considerations as to further assumptions of these methods. Looking solely at the probability prequisites for each method, we find that they differ markedly in their treatment of priors and conditionals. The modified binomial method relies on a single prior and not at all on conditionals. The catenating method relies on three separate priors and not all on conditionals. The decision tree method relies on a compound conditional but not at all on unconditional priors. The ranking method starts from a weakly-ordered set of priors to estimate conditionals. The cascaded inference method utilizes both priors and conditionals.

One important assumption concerns the independence of hypotheses -- are members of the set $\{H\}$ mutually exclusive or can they overlap? Along the same lines, are observations x of the set of evidence $\{X\}$ allowed to be partially or completely redundant, or must each observation be treated uniquely? The process by which each method proceeds is Bayesian with the notable exception of cascaded inference. (Further research is required as to how Baconian techniques may be brough to bear on the operation of the first five methods otherwise unmodified). At present, none of the methods handles the possibility of both unreliable data and unreliable behavior on the part of the examiner.

What is most interesting from the point of view of diagnosis is how each method enables one to evaluate the probative value of each piece of evidence -- that is, what term or expression (or change in terms or expressions) occurs at each step in the testing process such that the examiner sees how the last observation acquired has affected the

Figure 8
Comparison of probabilistic techniques:
prerequisites, processes, and outcomes

Technique:	SPRT	Modified Binomial	Catenating	Decision Tree	Ranking	Inexact Reasoning	Cascaded inference
Reference:	Wald	Sixtl	Choppin	Williams	Kmietowicz	Shortliffe	Schum
Prerequisites							
Unconditional priors							
$p(H_i)$	yes	yes	yes	no	} ranked	} approximated	yes
$p(H_j)$	no	no	yes	no			yes
$p(H_j)$	no	no	yes	no			no
Conditionals							
$p(x\|H_i)$	no	no	no	no	from table	approximated	yes
$p(x\|H_j)$	no	no	no	no			yes
$p(H\|x)$	no	no	no	no	no	yes	no
$p(x\|A,B...)$	no	no	no	yes	no	no	no
Independence							
$\{H\}$	yes	yes	yes	no	yes	?	no
$\{X\}$	yes	yes	yes	yes	no	no	no
Process							
Probability system	Bayesian	Bayesian	Bayesian	Bayesian	Bayesian	mixed Bayesian	Baconian
accounts for							
reliability	no	no	no	no	no	no	yes
guessing	no	yes	yes	no	no	no	?
Output	Λ_{ij}	Λ_{iH}	$P(H_{i,j})$	$P(H_{i,j})$	$P(\text{RANKING})$	MB_H, MD_H, CF_H	Λ_H
Stopping Rule	$\Lambda \geq$ upper $\leq$ lower	$\Lambda \geq$ upper $\leq$ lower	$P \geq$ constant	$P \geq$ constant	$P \geq$ constant	$CF \leq$ constant	$\Lambda \geq$ upper $\leq$ lower

relative standing of competing hypotheses. The output of the modified binomial technique is a set of likelihood coefficients Δ_H, one for each hypothesis. The output of the catenating Bayesian method is a set of probability estimates $p(H)$, one for each hypothesis, and a single term expressing the degree of uncertainty about their relative standings, SEF. The decision tree method outputs a probability estimate for every branch of the tree, allowing each hypothesis to be evaluated in context. The ranking method outputs a simple probability estimate for the entire set of hypotheses $P(\text{Rank})$, which shows the probability that the given ranking reflects the initial estimate of ranking of competing hypotheses. The inexact reasoning method outputs three separate terms per hypothesis at each step of the testing process; the last of these terms, CF_H, expresses the certainty with which the examiner can accept each hypothesis. The cascaded inference method outputs a likelihood coefficient for the hypotheses taken simply and taken jointly.

Four of the six methods are shown in Figure 9 as they step through a simulated testing session with very restrictive assumptions. For comparisons the results of the SPRT method are also shown. The examinee is presented only three choices for each of ten items; choice x_i is a reflection of hypothesis H_i, without guessing. A simulated testing session is used for which the examinee begins and ends with errors of type 1, but touches on other error types as well during the middle of the testing sequence; the examinee's response sequence is $\{1,2,3,3,2,1,2,1,1,1.\}$ Initial values were set at .6 for $p(x_i) \longrightarrow H_i$, and .2 for $p(X_i) \longrightarrow H_j$ α set at .25, β at .10. For illustrative purposes, computations are

Figure 9
Comparison of stopping using simulated data:

Step s		1	2	3	4	5	6	7	8	9	10
Response by examinee x		1	2	3	3	2	1	2	1	1	1
SPRT[1]	H_1	3.00	1.50	.75	.37	.19					
Λ_H	H_2	.50	1.50	.75	2.25	6.75 stop					
	H_3	.50	.25	.75	2.25	1.13					
Modified binomial[2]	H_1	2.20	1.76	1.41	1.13	0.90	1.98	1.59			
Λ_H	H_2	0.80	1.76	1.41	1.12	2.48	1.98	4.36 stop			
	H_3	0.80	1.76	1.41	3.10	2.48	1.98	1.59			
Catenating[3]	H_1	0.60	0.43	0.33	0.20	0.14	0.33	0.20	0.69	0.87 stop	
$p(H)$	H_2	0.20	0.43	0.33	0.20	0.43	0.33	0.60	0.23	0.10	
	H_3	0.20	0.14	0.33	0.68	0.43	0.33	0.20	0.07	0.03	
S.E.F.[4]		0.41	0.43	0.48	0.47	0.43	0.48	0.43	0.34	0.20	
Inexact reasoning[5]	H_1	.41	.01	-.23	-.37	-.48	-.22	-.27	-.23	-.05	0
CF_H	H_2	-.40	-.40	.01	-.23	-.37	-.13	-.22	-.08	-.17	-.18
	H_3	-.40	-.64	.23	-.01	-.13	-.27	-.27	-.31	-.33	-.34
Ranking[6] Weak 1≥2≥3		n/a[7]	n/a	n/a	0.20	0.15	n/a	0.39	1.00 stop		
p(RANK) Strong 1>2>3		n/a	n/a	n/a	0.18	0.18	n/a	0.15	0.39	1.00 stop	

1 Wald (1947). See formula (1).
2 Sixtl (1974). See formula (2).
3 Choppin in McArthur and Choppin (1983). See formula (3).
4 Shanon entropy function (Gleser & Collen, 1972). See formula (4).
5 Shortliffe and Buchanan (1975). See formula (5).
6 Kmietowicz & Pearson (1981); Horbar (1983). See formula (6).
7 Not appropriate to calculate at this step.

shown for the unmodified sequential probability ratio test, which concludes step 5 with support for H_2 . The modified binomial method concludes at step 7 with support for H_2 . The catenating Bayesian method concludes at step 10 with support for H_1 . The ranking method (shown using an estimated order of $H_1 > H_2 > H_3$) concludes at step 8 if the ranking is assumed to be weak, step 10 if strong. The inexact reasoning method fails to conclude by step 10. (Because the remaining two methods, decision path and cascaded inference, require many further initial assumptions, they are not included in this illustration).

Conclusion

That the separate techniques fail to agree on where to stop and which competing hypothesis to support comes as no surprise. There are numerous reasons why agreement between techniques is unlikely. The initial statistical prerequisites are numerous, and unevenly taken into account. Unconditional priors do not have the same effect as simple conditionals or compound conditionals. The inclusion of each new term predictably affects computations, such that in general, with all else held the same, the larger number of priors and conditionals the longer it will take to reach the stopping rule. Further complications are added if members of $\{H\}$ or $\{X\}$ are not independent, are not unambiguous or not properly targeted to the test, and so forth.

Fischhoff and Beyth-Marom (1983) offer an extensive list of pitfalls of hypothesis evaluation:

- untestable hypotheses (absent, nonevaluatable, too complex, nonexclusive)
- wrong component probabilities (misrepresented. miscalibrated, nonconforming)

- wrong prior probabilities (incomplete, fallacious, unrepresentative)
- wrong likelihood ratios (distorted, H_{ij} neglected, non-causal)
- incorrect aggregation (rules misapplied, values computed extraneously)
- inadequate search of evidence (questions non-diagnostic, inefficient, incomplete)
- uncertain consequences (inadequate opportunity or resources to pursue optimal cause of action)

In particular, a problem that confronts a diagnostician after assessing the available evidence from a test is how to convert such knowledge into concrete actions.

> "...Knowledge of the possible actions is essential in determining what information to gather. Two...judges who contemplated different actions, or evaluated their consequences differently, might justifiably formulate different hypotheses and collect different data even though they agreed on the interpretation of all possible data (Fishhoff & Beyth-Marom, 1983, p.250).

None of the techniques portrayed here succeed in addressing all of their concerns.

Sequence considerations, which contribute to the nonindependence of $\{X\}$, are taken into account only by the more complicated methods. None explicitly treats the complex relationship between an examinee's ability, likelihood of guessing, and performance. None explicitly indicates to which next item is optimal -- that is, optimality of branching continues to depend on how close the members of $\{H\}$ are to one another, how rapidly the examiner will like to converge on a single H, and how exhaustive a search of $\binom{m}{n}$ combinations is desired. A very fast sequence can be derived if one steps through a selection of $\{H\}$ for which all but one are known to be exceedingly unlikely for the examinee. The same is true if one chooses liberal values for α and β, or shapes the stopping rule to favor an otherwise inconclusive outcome.

Bayesian analysis is only one of several systems which treat probabilistic data. However, it has been the overwhelming system of choice despite repeated objection if only because completely explicated alternatives are rare. A system which allowed incompleteness, nonmultiplicative joint probabilities, and conditional nonindependence of H would be preferable in the context of diagnostic testing[1]. The Baconian system of Cohen appears to meet these needs. For example, Cohen's system does not include mathematical additivity, an inherent property of Bayesian techniques, so $P(H_i) = 0$ does not mean that $P(H_j) = 1$. Conjunction of probabilities, which is multiplicative in Bayesian analysis, is handled by taking the minimum $p(H) = P(H_1 \cap H_2 \cap \ldots) = \min(P(H_k))$.

Remaining for further study is how the rules of Baconian probability manipulations might apply to the Bayesian techniques presented here. A closely related issue is whether the Baconian system is as sensitive to the choice of prior probabilities as the exact Bayesian systems which are shown above.

A further set of issues about statistics for diagnostic testing concerns a facet of test design mentioned only fleetingly in this paper: the relations of items to ability θ. Indeed, only if the hypotheses are well-bounded and the choices for test items are demonstrably associated

---

1 Incompleteness: $P(E|F+) = 0$ and $P(E|F-) = 0$ are allowed; in Bayesian analysis if $P(E|F+) = 1$, $P(E|F-)$ must equal 0.

Nonmultiplicative joint probabilities: the joint occurence of two relatively rare events need not be less than their separate occurence; in Bayesian analysis, soon enough the multiplicative rule leaves any hypothesis p(H) supported less than $p = .5$.

Conditional nonindependence of $\{H\}$: hypotheses may be evaluated even if they overlap, or incompletely requested at each stage in a hierarchy of hypotheses.

with those hypotheses will a diagnostic inference system succeed. That is, if the hypotheses available for assessment are unproductive (ill-suited, poorly framed, highly redundant, or otherwise off target), no amount of statistical manipulation will rescue the examiner from a possibly erroneous and certainly frustrating conclusion. Likewise, if the choices available to an examinee are poor reflections of good hypotheses, the examiner will also experience no closure at all, or potential diagnostic inaccuracies if closure is reached.

REFERENCES

Adams, J.B. (1976). A probability model of medical reasoning and the MYCIN model. Mathematical biosciences, 32, 177-186.

Cohen, L.J. (1977). The probable and the provable. London: Oxford University Press.

Cohen, L.J., & Hesse, M. (Eds.). (1980). Applications of inductive logic. Oxford: Caredon Press.

Ferguson, R.L. (1969). Computer-assisted criterion-referenced measurement. University of Pittsburgh, Learning Research and Development Center, Working Paper #49.

Ferguson, R.L., & Novick, M.R. (1973). Implementation of a Bayesian system for decision analysis in a program of individually prescribed instruction. Iowa City: American College Testing Program Research Report #60.

Fink, D.J., & Galen, R.S. (1981). Probabilistic approach to clinical decision support. In B.T. Williams (Ed.), Computer aids to clinical decisions. Boca Raton, FL: CRC Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. Psychological Review, 90, 239-260.

Gleser, N.A., & Collen, M.F. (1972). Towards automated medical decisions. Computers and biomedical research, 5, 180-189.

Gorry, G.A., & Barnett, G.O. (1968). Experience with a model of sequential diagnosis. Computers and biomedical research, 1, 490-507.

Hobar, J.D. (1983). Raising ranked probabilities: A Bayesian approach to incomplete knowledge. Computer and Biomedical Research, 16, 367-377.

Kmietowicz, Z.W., & Pearson, A.D. (1981). Decision theory and incomplete knowledge. Hampshire, England: Gower.

McArthur, D.L., & Choppin, B.H. (1983). Evaluation of diagnostic hypotheses. Final Report NIE Grant G-83-0001. Los Angeles: Center for the Study of Evaluation.

McArthur, D.S., & Roberts, G. (1982). Manual for the Roberts Apprception Test for Children. Los Angeles; Western Psychological Services.

Pressey, S.L. (1926). A simple apparatus which gives tests and scores -- and teaches. School and Society, 23, 373-376.

Roid, G.H., & Haladyna, T.M. (1982). A technology for test-item writing. New York: Academic Press.

Schum, D.A. (1979). A review of a case against Blaise Pascal and his heirs. Michigan Law Review, 77, 446-483.

Schum, D.A. (1977). The behavioral richness of cascaded inference models: Examples in jurisprudence. In N.J. Castella, D.B. Pisoni, and G.R. Potts (Eds.). Cognitive theory, Hillsdale N.J.: Erlbaum, Volume 2, 149-174.

Schum, D.A., & Kelly, C.W. (1973). A problem in cascaded inference: Determining the inferential impact of confirming and conflicting reports from several unreliable sources. Organizational behavior and human performance, 10, 404-423.

Schum, D.A., & Martin, A.W. (1980). Empirical studies of cascaded inferences in jurisprudence: Methodological considerations. Rice University Department of Psychology Research Report 80-01.

Shortliffe, E.H., & Buchanan, B.G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23, 351-379.